DOCUMENT RESUME

ED 231 884 TM 830 501

AUTHOR Johnston, Shirley H.; And Others
TITLE The Effects of Violating the Beta-Binomial Assumption on Huynh's Estimates of Decision Consistency for Mastery Tests.
PUB DATE Apr 83
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bias; Criterion Referenced Tests; Data; *Mastery Tests; *Mathematical Models; Reliability; Research Design; *Sample Size; *Scores; Statistics
IDENTIFIERS *Beta Binomial Test Model; Computer Simulation; *Estimates of Decision Consistency (Huynh); Repeated Testing; Test Equivalence; Test Length

ABSTRACT

A computer simulation was undertaken to determine the effects of using Huynh's single-administration estimates of the decision consistency indices for agreement and coefficient kappa, under conditions that violated the beta-binomial assumption. Included in the investigation were two unimodal score distributions that fit the model and two bimodal types that did not. While varying test length, mastery level, and sample size, estimates were calculated for both indices from two data sets, simulated to represent scores on equivalent forms. Then Huynh's method was applied to one of the sets to obtain the beta-binomial estimates. Statistics were computed to measure the magnitude and direction of bias resulting from the application of Huynh's technique. The results of the study did not provide conclusive evidence that Huynh's method is adversely affected by violation of the beta-binomial assumption. There was, however, clear indication that the estimates are strongly influenced by certain configurations of scores. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. H. Johnston

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED231884

The Effects of Violating the Beta-Binomial Assumption on Huynh's Estimates of Decision Consistency for Mastery Tests

Shirley H. Johnston, Mary A. MacDougall, and Michael Caldwell

University of Virginia

Decision consistency has been widely endorsed as a criterion-referenced approach to reliability. When a test is designed to separate masters from nonmasters, the proportion of examinees consistently classified through repeated testing, with either the same test or equivalent forms, can be quantified by the agreement index, p. Because the size of this measure is affected substantially by the mastery-nonmastery composition of the group being tested, Cohen's (1960) corrected-for-chance coefficient kappa (κ) is also strongly recommended. Both are readily computed from the results of two testings but can be estimated, according to Huynh (1976), on the basis of a single test.

Huynh's procedure is formulated on the beta-binomial test model described by Keats and Lord (1962). This implies that errors of measurement

---

• The first author is an Associate Professor of Mathematics at J. Sargeant Reynolds Community College in Richmond, Virginia. The research reported in this paper was conducted at the University of Virginia in 1981 and was submitted as a doctoral dissertation to the faculty of the School of Education, Department of Research and Evaluation.

Printed in U.S.A.

TM 830 301

can be represented by the discrete binomial error model (Lord & Novick, 1968) and that true scores are distributed within the population as a member of the family of probability densities called beta. Given observed scores from one administration of a mastery test, Huynh's method yields estimates of the parameters of the beta distribution of true scores. The same parameters are expected to characterize the joint distribution of scores on two random variables, or two parallel forms of the test. Therefore, a joint density function, based on these parameters, can be used to calculate the probability that any individual would receive scores that are either above the mastery level on both tests or below the mastery level on both tests. This probability is then taken to be an estimate of the proportion of consistent mastery-nonmastery classification decisions, or the raw agreement index, p. A correction for chance yields the corresponding estimate for kappa. Huynh provides computational formulas for both p and κ (1976, pp. 254-255), as well as a computer program (1980, Appendix B) and tables for use with short tests (1980, Appendix A). For persons who have no access to a computer, approximate procedures are described by Huynh (1976) and by Peng and Subkoviak (1980).

The convenience of a single test administration and the availability of computational aids contribute to the popularity of Huynh's technique. It will undoubtedly be applied in situations in which its basic assumptions are not met. In particular, the beta form for true scores does not allow for multimodal data, a possible outcome when mastery tests are given. Studies by Gross and Shulman (1980) and by Huynh and Saunders (1980), using typical criterion-referenced test data, indicate that violation of the beta-binomial assumption will not seriously affect results. Neither of

these investigations, however, included multimodal distributions of test scores, a condition identified by Marshall and Serlin (1979) as a matter of some concern.

The purpose of this study was to compare Huynh's single-administration estimates of decision consistency with the estimates obtained from repeated testing, when the beta-binomial assumption failed to hold. Emphasis was on the beta representation for true scores, rather than on the binomial error model, as in the previously cited research.

Method

Test score data were generated by computer to simulate sampling from four selected populations, each of which was considered likely to appear in practice. The Kolmogorov-Smirnov one-sample test (Siegel, 1956) was used to assess the goodness of fit between an exemplary set of scores and the theoretical beta-binomial distribution. Unimodal data sets A and B conformed to the model, while bimodal C and D did not (see Figure 1). It should be noted that bimodality alone does not guarantee a lack of fit and that the score distributions used in this study were carefully constructed to achieve this result.

Test lengths were taken to be n = 5, 10, 20, and 50 items. Criterion-referenced tests are usually short, reflecting a single educational objective and containing from 3 to 20 items (Popham, 1980). However, in certain applications, such as minimum competency testing, it is common practice to combine the subscales for several objectives and to award a composite score. For this reason, 50-item tests were included in the study.

Mastery levels were set at 60% and 80% for the 5-item tests and 60%, 70%, and 80% for all others. Sample sizes were chosen as M = 30, to typify classroom size, and M = 300 for a larger size.

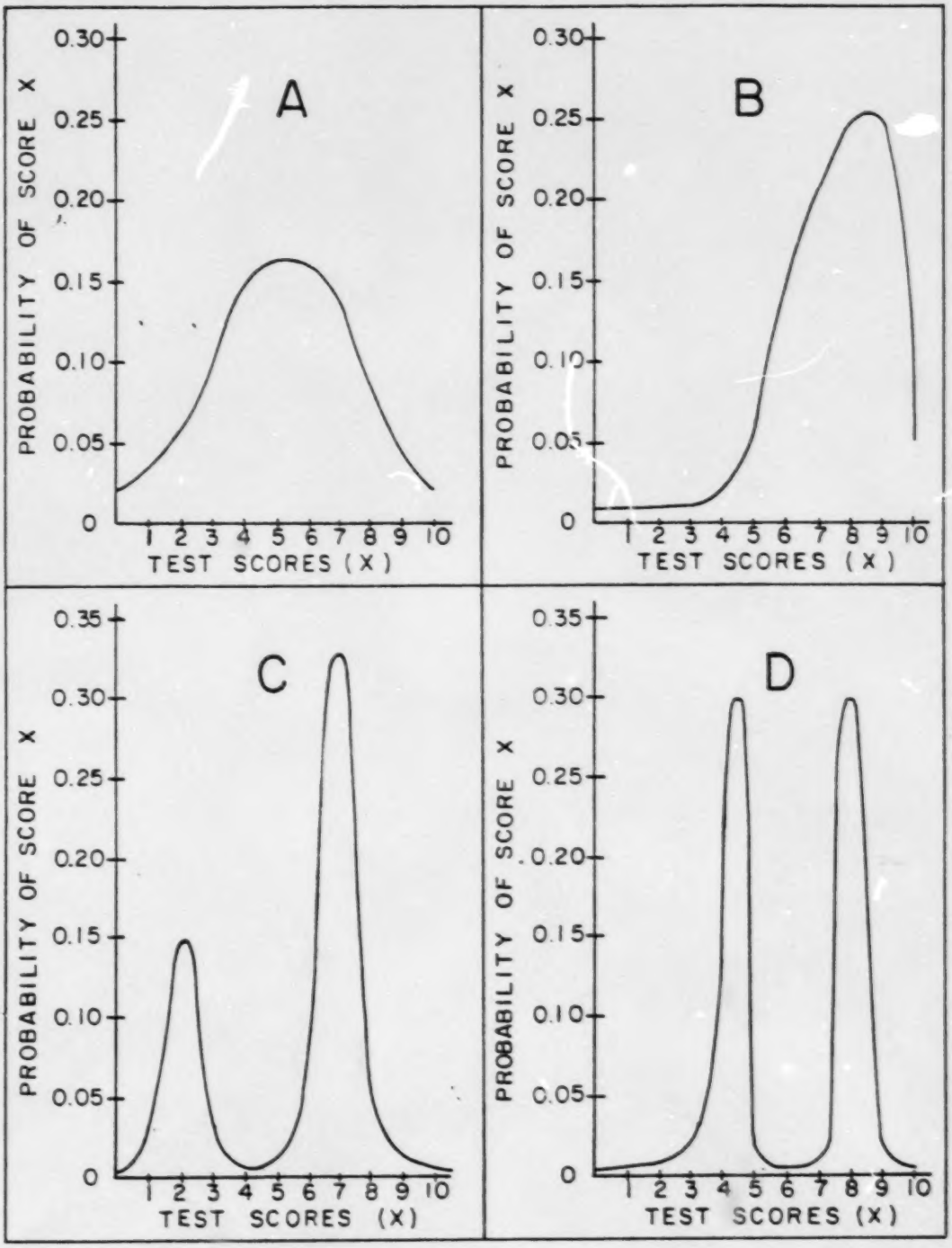

Figure 1

Theoretical distributions of true scores for tests of length n = 10 and M = 300 examinees

A computer program written by the author was used to obtain data and to perform all computations. Subroutines from the International Mathematical and Statistical Libraries (IMSL) and from Huynh (1980, Appendix B) were incorporated into the program to generate scores and to calculate the beta-binomial estimates for p and κ.

The study required that each examinee be assigned two scores to simulate the result of testing with equivalent forms. To achieve this, each set of data produced by the IMSL subroutine, GGVCR, was assumed to represent a sample of true scores from the parent population (assigned score distribution A, B, C, or D). Observed scores for each of the two forms were computed from the true scores by adding a random error component within the range of ±20%.

Because true scores and random errors were produced from continuous distributions, calculated observed scores were then rounded to the nearest integer value. If this resulted in any out-of-range scores ($x < 0$ or $x > n$), they were taken to be zero or n.

The pairs of observed scores derived for each subject and the computational formulas given by Huynh (1976) were used to calculate the dual-administration coefficients, $\hat{p}_R$ and $\hat{\kappa}_R$. Huynh's method was then applied to the first set of observed scores only, to arrive at the single-administration estimates, $\hat{p}_B$ and $\hat{\kappa}_B$, based on the beta-binomial model. Finally, the deviations $(\hat{p}_B - \hat{p}_R)$ and $(\hat{\kappa}_B - \hat{\kappa}_R)$ were obtained.

The simulation consisted of 500 replications for the large samples ($M = 300$) and 1000 for the classroom-size samples ($M = 30$) because they were expected to be less stable.

It has been established (Huynh, 1980) that the estimate of the raw agreement index p, calculated from data on equivalent test forms, is an

unbiased estimate of the proportion of consistent classifications in the population. Similarly, the dual-administration estimate of kappa is an asymptotically ($M \to \infty$) unbiased estimate of the proportion of consistent classifications corrected for chance. Although the population values for p and κ were unknown at the beginning of the simulation, the averages over all replications of $\hat{p}_R$ and $\hat{\kappa}_R$ can be considered close approximations. Therefore, the statistic of interest was the bias of Huynh's estimates of p and κ, as measured by the averages of the deviations $(\hat{p}_B - \hat{p}_R)$ and $(\hat{\kappa}_B - \hat{\kappa}_R)$. For the 88 crossed cases of distribution shape, test length, mastery level, and sample size, the magnitude, direction, and variability of bias were tabulated and compared.

Results

A particular feature of the design of this study was that the shapes of the hypothetical test score distributions could be held constant while other conditions were varied. For example, simulated data sets could be considered to have come from tests that differed in length but were otherwise essentially the same. (This is not what would be expected if individual tests were lengthened by adding comparable items.) The adequacy of the IMSL subroutine GGVCR in generating true scores from the specified distribution functions was confirmed by the similarity of means and standard deviations for small and large samples and by the proportionality of both statistics for proportional values of n.

Values of $\hat{p}_R$ and the biases of their beta-binomial estimates are displayed, for the large-size samples, in Table 1. Biases ranged from -.30 to +.23, with both extremes recorded for distribution type C which had two unequal and widely separated modes. The same distribution produced the most

Table 1

Mean Bias of $\hat{P}_B$ in Estimating $\hat{P}_R$

$\underline{M}$ = 300

n	c	Distribution A		B		C		D	
		$\hat{P}_R$	Bias	$\hat{P}_R$	Bias	$\hat{P}_R$	Bias	$\hat{P}_R$	Bias
5	3	.782	-.198	.863	-.143	.955	-.299	.813	-.184
	4	.840	-.135	.716	-.179	.679	.013	.807	-.194
10	6	.789	-.049	.819	-.048	.865	-.090	.883	-.144
	7	.817	-.045	.737	-.030	.704	.080	.877	-.152
	8	.867	-.038	.700	-.010	.694	.117	.770	-.008
20	12	.793	.034	.798	.035	.869	.034	.916	-.097
	14	.828	.023	.722	.076	.686	.169	.842	-.023
	16	.881	.008	.706	.087	.719	.157	.757	.084
50	30	.795	.099	.785	.108	.787	.120	.924	-.036
	35	.833	.075	.715	.159	.681	.231	.819	.067
	40	.889	.044	.713	.160	.737	.186	.754	.147

positive bias overall. On the other hand, there was likely to be more negative bias in the case of the other bimodal example, D. For all four distributions, there was a trend from negative to positive bias as the number of items increased. Distance between the cutoff score and the mean yielded mixed results throughout.

For kappa, biases followed the same patterns as those for p, but were approximately twice as large. This is consistent with the findings of Huynh and Saunders (1980), who attribute it to a factor in the denominator of the equation defining κ. The bias of $\hat{\kappa}_B$ in estimating $\hat{\kappa}_R$ for large samples is shown in Figure 2.

Standard errors computed for the estimates of p_R and the indicators of bias, $(\hat{p}_B - \hat{p}_R)$, revealed that all measures were fairly stable, regardless of test length, cutoff score, or distribution type. Corresponding statistics for kappa were more variable, as the literature had indicated they would be.

Outcomes of the simulation for small samples were essentially the same. The estimates of p_R and κ_R for each case were very close to those for their large-sample counterparts. Measures of bias were also similar but exhibited standard errors approximately three times as large.

Discussion

The results of this study did not clearly indicate that Huynh's single-administration estimates of decision consistency are more biased for bimodal distributions of scores. In particular, the indices did not behave in the same way for both bimodal types C and D. If it could be argued that underestimation is more desirable than overestimation, then the method frequently worked better for distribution D than it did for either of the two unimodal data sets that conformed to the theoretical model. There was also no particular number of items for which violating the distributional assumption

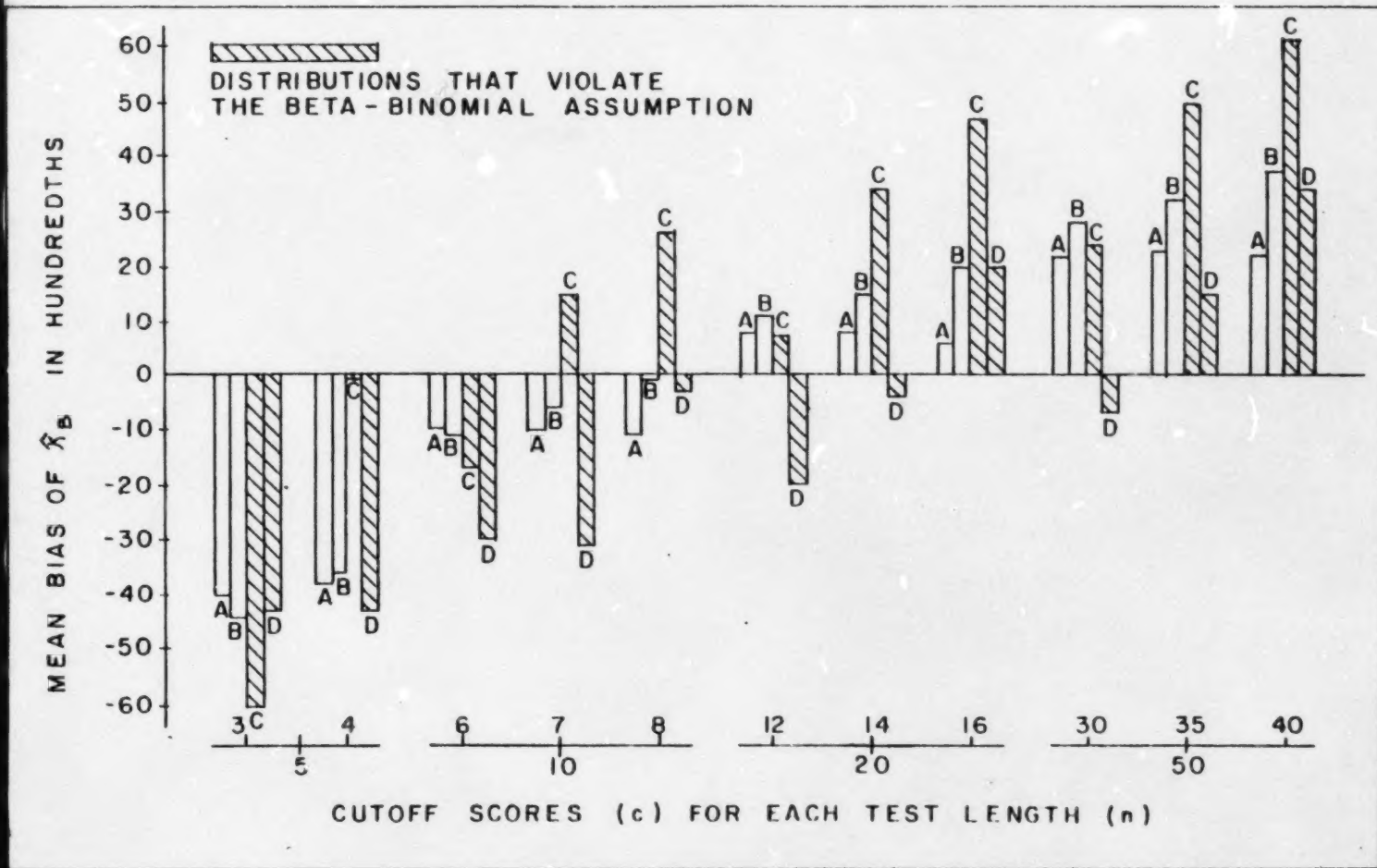

Figure 2

Mean bias of Huynh's single-administration index $\hat{\kappa}_B$ in estimating $\hat{\kappa}_R$ for $\underline{M}$ = 300

produced a noticeable effect. A moderate difference in bias between the bimodal and unimodal types was observed as the position of the cutoff score was varied with respect to the mean.

The systematic change in bias from negative to positive as test lengths were increased was the most compelling outcome of the study. Just as traditional measures of reliability increase when a test is lengthened, so do the agreement coefficients p and κ. This is true whether they are computed from two tests or estimated using the beta-binomial model. However, in this experiment, distributional shapes were held constant with means and standard deviations in proportion for tests of varying length. As a result, the computed values of p_R and κ_R showed little or no change and bias steadily increased.

The implication for those who wish to use Huynh's method is that, if it works well for shorter tests and if these are lengthened by adding comparable items, the effects noticed in this investigation would not occur. On the other hand, if a 50-item test produced a score distribution like any of the four used in the study, then Huynh's estimates would be highly inflated. This would also be true if a 20-item test yielded a bimodal configuration like C. Likewise, p and κ would be considerably underestimated for a 5-item test when any of the four given distributions occurred and moderately so with $n = 10$ for a bimodal data set like D.

Clearly, the influence of distributional shape on Huynh's estimates of p and κ, as well as on the indices themselves, cannot be ignored. Practitioners should be wary of using the beta-binomial estimates for tests of more than 20 items and expect that for very short tests the coefficients might be spuriously low. Finally, classroom teachers should be warned that

not only were all of the previously stated effects present for samples of size 30 but considerable fluctuation in the measures was also observed.

References

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Gross, A.L., & Shulman, V. The applicability of the beta-binomial model for criterion-referenced testing. Journal of Educational Measurement, 1980, 17, 195-201.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. Assessing efficiency of decisions in mastery testing. In H. Huynh & J.C. Saunders (Eds.), Solutions for some technical problems in domain-referenced mastery testing (Final Report). Columbia, SC: University of South Carolina, 1980.

Huynh, H., & Saunders, J.C. Accuracy of two procedures for estimating reliability of mastery tests. Journal of Educational Measurement, 1980, 17, 351-358.

IMSL Library. Houston: International Mathematical and Statistical Libraries, 1980.

Keats, J.A., & Lord, F.M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Marshall, J.L., & Serlin, R.C. Characteristics of four mastery test reliability indices: Influence of distribution shape and cutting score. Paper presented at the annual meeting of the American Educational Research Association, April 1979, San Francisco. (ERIC Document Reproduction Service No. ED 171 778)

Peng, C.J., & Subkoviak, M.J. A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. Journal of Educational Measurement, 1980, 17, 359-368.

Popham, W.J. Domain specification strategies. In R.A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

TM 830 501

ABSTRACT

A computer simulation was undertaken to determine the effects of using Huynh's single-administration estimates of the decision consistency indices, $\underline{p}$ and kappa, under conditions that violated the beta-binomial assumption. Included in the investigation were two unimodal score distributions that fit the model and two bimodal types that did not. While varying test length, mastery level, and sample size, estimates were calculated for $\underline{p}$ and kappa from two data sets, simulated to represent scores on equivalent forms. Then Huynh's method was applied to one of the sets to obtain the beta-binomial estimates. Statistics were computed to measure the magnitude and direction of bias resulting from the application of Huynh's technique.

The results of the study did not provide conclusive evidence that Huynh's method is adversely affected by violation of the beta-binomial assumption. There was, however, clear indication that the estimates are most accurate for 10-item tests and that they are strongly influenced by certain configurations of scores.